Tratamiento Masivo de Datos Utilizando Técnicas de Machine Learning

Claudia Russo[1], Hugo Ramón[1], Nicolás Alonso[1], Benjamin Cicerchia[2], Leonardo Esnaola[1], Juan Pablo Tessore[2]

[1]Instituto de Investigación y Transferencia en Tecnología (ITT) / Escuela de Tecnología/ Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA)
Sarmiento y Newbery, 236-4636945/44

[2]Becario de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC)

claudia.russo@itt.unnoba.edu.ar / hugo.ramon@itt.unnoba.edu.ar / nicolas.alonso@itt.unnoba.edu.ar / lucas.cicerchia@itt.unnoba.edu.ar / leonardo.esnaola@itt.unnoba.edu.ar / juanpablo.tessore@itt.unnoba.edu.ar

Resumen

Machine Learning es un área de la inteligencia artificial que engloba un conjunto de técnicas que hacen posible el aprendizaje automático a través del entrenamiento con grandes volúmenes de datos. Hoy en día existen diferentes modelos que utilizan esta técnica y consiguen una precisión incluso superior a la de los humanos en las mismas tareas, por ejemplo en el reconocimiento de objetos en una imagen.

La construcción de modelos de *Machine Learning* requiere adaptaciones propias debido a la naturaleza de los datos o a la problemática a la que se aplica. Así, surge la necesidad de investigar las diferentes técnicas que permitan obtener resultados precisos y confiables en un tiempo razonable.

Palabras clave:

Machine Learning, Big Data, Sistemas Inteligentes.

Contexto

Esta línea de investigación forma parte del proyecto "Tecnologías exponenciales en contextos de realidades mixtas e interfaces avanzadas." aprobado por la Secretaría de Investigación, Desarrollo y Transferencia de la UNNOBA en el marco de la convocatoria a Subsidios de Investigación Bianuales (SIB2015). A su vez se enmarca en el contexto de planes de trabajo aprobados por la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires y por la Secretaría de Investigación de la UNNOBA en el marco de la convocatoria "Becas de Estudio Cofinanciadas 2015 CIC Universidades del interior bonaerense".

El proyecto se desarrolla en el Instituto de Investigación en Tecnologías y Transferencia (ITT) dependiente de la

mencionada Secretaría, y se trabaja en conjunto con la Escuela de Tecnología de la UNNOBA.

El equipo está constituido por docentes e investigadores pertenecientes al ITT y a otros Institutos de Investigación, así como también, estudiantes de las carreras de Informática de la Escuela de Tecnología de la UNNOBA.

Introducción

Desde que las primeras computadoras programables fueron concebidas, las personas se preguntaron si tendrían la capacidad de pensar, de aprender y de convertirse en "máquinas inteligentes".

El campo de la ciencia que se encarga de resolver este interrogante se denomina inteligencia artificial. Se trata de un área multidisciplinaria, que a través de ciencias como las ciencias de la computación, la matemática, la lógica y la filosofía, estudia la creación y diseño de sistemas capaces de resolver problemas cotidianos por sí mismos, utilizando como paradigma la inteligencia humana [1]. Para que una maquina pueda comportarse de manera inteligente debería ser capaz de resolver problemas de la manera en que lo hacen los humanos, es decir, en base a la experiencia y el conocimiento [2]. Esto implica que debería ser capaz de modificar su comportamiento en base a cuan precisos son los resultados obtenidos comparados con los esperados.

En este sentido podemos encontrar tres grandes grupos de algoritmos de *Machine Learning* [3]:

- Algoritmos supervisados: estos algoritmos utilizan un conjunto de datos de entrenamiento etiquetados (preclasificados), los cuales procesan para realizar predicciones sobre los mismos, corrigiéndolas cuando son incorrectas. El proceso de entrenamiento continúa hasta que el modelo alcanza un nivel deseado de precisión.
- Algoritmos semi-supervisados: combinan tanto datos etiquetados como no etiquetados para generar una función deseada o clasificador. Este tipo de modelos deben aprender las estructuras para organizar los datos así como también realizar predicciones.
- Algoritmos no supervisados: El conjunto de datos no se encuentra etiquetado y no se tiene un resultado conocido. Por ello deben deducir las estructuras presentes en los datos de entrada, lo puede conseguir a través de un proceso matemático para reducir la redundancia sistemáticamente u organizando los datos por similitud.

Dentro de esta clasificación podemos además encontrar un gran número de algoritmos específicos con diferentes características para el tratamiento de los datos. Entre los más relevantes encontramos:

- *Deep Learning*: consiste en la utilización de algoritmos para hacer representaciones abstractas de la información y facilitar el aprendizaje automático [4].
- *Active Learning*: es un caso especial de aprendizaje semi-supervisado donde el algoritmo de aprendizaje

puede interactuar con un usuario u otra fuente de información para obtener los resultados deseados [5].

- *Support Vector Machines*: busca la maximización de la distancia entre la recta o el plano y las muestras que se encuentran a un lado u otro. En el caso que las muestras no sean linealmente separables se utiliza una transformación llamada *kernel* [6] [7].

Líneas de Investigación, Desarrollo e Innovación

La presente investigación se encuadra dentro del eje "Tratamiento masivo de datos" y su procesamiento a través de sistemas inteligentes. En este sentido se pretende procesar señales provenientes de fuentes diversas, según la problemática investigada, para construir los conjuntos de entrenamiento necesarios. Así como también, la selección, diseño y desarrollo de un modelo que utilice alguno de los algoritmos relevados para lograr una correcta clasificación y predicción.

Se deberán abarcar las siguientes cuestiones:

- Obtención de un conjunto de datos suficientemente representativo para la problemática que se desea abordar y su clasificación.
- Pre procesamiento de las señales para lograr su normalización y adecuación.
- Resolver cuestiones relacionadas con el procesamiento de datos en tiempo real y optimización de su funcionamiento.
- Análisis y selección de los distintos algoritmos de *Machine Learning* apropiados para el tipo de señal recibida (imágenes, video, sonido o incluso texto) y para la problemática de su aplicación.
- Evaluación de fiabilidad y desempeño de las diferentes técnicas de *Machine Learning* aplicadas.

Resultados y Objetivos

Se espera que la presente línea de I/D permita adquirir conocimientos específicos sobre las diferentes técnicas de *Machine Learning*, con el propósito de desarrollar modelos capaces de predecir y clasificar las señales involucradas en la problemática que se intenta resolver, obteniendo un comportamiento inteligente de manera automática.

Debido a que la universidad se encuentra dentro de la pampa húmeda, una de las regiones más relevantes en lo que respecta a producción agrícola, se pretende combinar agricultura de precisión con técnicas de machine learning y remote sensing [8] con el objetivo de dar soporte a la toma de decisiones en el sector.

Por otro lado, debido a las necesidades de los municipios de la región, se vislumbra la posibilidad de trabajar en la prevención y la detección de diferentes tipos de comportamiento, problemática que también puede ser atacada con este tipo de técnicas.

También se prevé la aplicación de machine learning en el análisis del texto en publicaciones periodísticas, contenido en foros y redes sociales, con la finalidad de encontrar patrones dentro de esos datos que permitan predecir comportamientos futuros en ámbitos específicos.

Así mismo, se busca generar informes técnicos en base al trabajo realizado, en donde se registren los avances, el grado de implementación y los resultados obtenidos. Como así también difundir y transferir los resultados y logros alcanzados mediante la presentación y participación en diferentes congresos, jornadas y workshops de carácter nacional e internacional vinculados a la temática de estudio.

Formación de Recursos Humanos

En esta línea de I/D se han obtenido y se encuentran desarrollando actualmente dos Becas de Estudio Cofinanciadas otorgadas por la Comisión de Investigaciones Científicas (CIC) y la UNNOBA. Asimismo se espera desarrollar cuatro tesis doctorales y dos tesinas de grado, dirigidas por miembros de este proyecto.

Bibliografía

[1] Assessment of the Commercial Applicability of Artificial Intelligence in Electronic Businesses. Thomas Kramer. Diplom.de. 2002.

[2] Data Classification Algorithms and Applications, Charu C. Aggarwal, CRC Press, 2015.

[3] Machine Learning An Algorithmic Perspective Second Edition, Stephen Marsland, CRC Press, 2015.

[4] A Deep Learning. Book in preparation for MIT Press. Bengio, Y., Goodfellow, I. and Courville, USA, 2015.

[5] Active Learning Literature Survey, Settles Burr, Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 2014.

[6] A Tutorial on Support Vector Machines for Pattern Recognition, Christopher J.C. Burges, Kluwer Academic Publishers, 1998.

[7] Top 10 algorithms in data mining, Xindong Wu et al. Knowledge and Information Systems 2008.

[8] Machine learning in remote sensing data processing, Gustavo Camps-Valls, IEEE International Workshop on Machine Learning for Signal Processing, 2009.